academicJournals

Vol. 13(41), pp. 4052-4060, 8 October, 2014
DOI: 10.5897/AJB2014.14218
Article Number: 2506B1B47973
ISSN 1684-5315

http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

Targeted parallel sequencing of the *Musa* species: Searching for an alternative model system for polyploidy studies

Ude, George N[1]*, Acquaah, George[1], Irish, Brian M[2] and Das, Aditi[1]

[1]Department of Natural Sciences, Bowie State University, Bowie, Maryland, United States.
[2]USDA-ARS, Tropical Agriculture Research Station, Mayaguez, Puerto Rico.

Received 30 September, 2014; Accepted 3 October, 2014

Modern day genomics holds the promise of solving the complexities of basic plant sciences, and of catalyzing practical advances in plant breeding. While contiguous, "base perfect" deep sequencing is a key module of any genome project; recent advances in parallel next generation sequencing (NGS) technologies has opened up new avenues for answering biological questions in moderate to large genomes of complex polyploid species like banana. Most edible cultivated bananas belong to the Eumusa section of the Musaceae, and are diploid or triploid hybrids from their wild diploid ancestors: *Musa acuminata* (A-genome) alone or from hybridization with *Musa balbisiana* (B-genome). In this study, a second-generation parallel sequencing method was implemented to identify nucleotide variants in *Musa* spp. This strategy reduced genome complexity by enrichment with a hybridization capture library, targeting primarily exons of coding genes. The resulting marker dataset was successful in sampling broadly within the A and B genome groups and their derived hybrids. The study confirms the sequence diversity of *Musa* on a genome-wide scale even in a modest subset of *Musa* cultivars. Importantly, the experimental approach undertaken here is an efficient means of producing data for the design of high and low-density nucleotide polymorphism (single-base substitutions, small insertions and deletions or INDELs) genotyping assays applicable to a wide range of *Musa* cultivars. Thus, an excellent alternative method is reported, for characterizing associations between genotypic and phenotypic variation in *Musa* by using sequence variants as molecular markers.

Key words: Sequence capture, in-solution hybridization, nucleotide variants, polyploidy, bananas (*Musa* spp.)

INTRODUCTION

Bananas (*Musa* spp.) are very important in the diets of people and national economies in the tropics and subtropics. Global production of bananas and plantains (a type of banana) was estimated at about 140 million

*Corresponding author. E-mail: gude@bowiestate.edu.

Abbreviations: NGS, Next generation sequencing, **MNPs,** multi-nucleotide polymorphisms; **INDELs,** insertions and deletions; **MQ,** mapping quality; **CDS,** coding sequence; **SNPs,** single nucleotide polymorphisms; **AFLP,** amplified fragment length polymorphism.

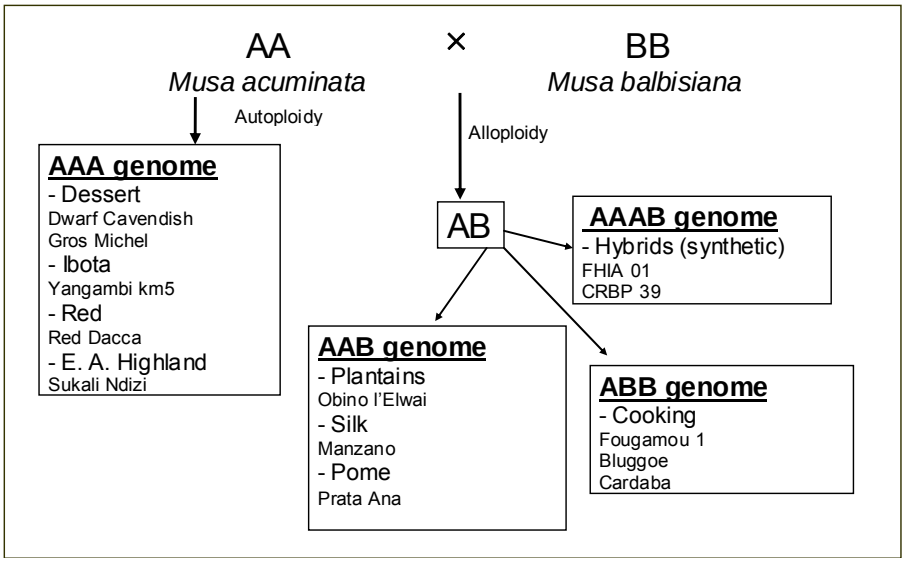

Figure 1. Schematic illustration of how a representative group of modern day edible polyploid *Musa* spp. sub-groups cultivars originated.

metric tons in 2012 (FAOSTAT, 2014). The crop is vegetatively propagated using land races and a few recently developed modern cultivars. As a consequence of its reproductive system, banana has a narrow genetic base, making it highly susceptible to diseases, pests and abiotic stresses in the environment. To genetically improve the crop, breeders depend on genetic recombination of improved diploids, crossed into seed-setting triploids to produce sterile polyploid cultivars. To facilitate modern banana improvement, breeders could benefit from an understanding of the phylogenic ancestry of the crop as well as pathways of domestication of the major cultivar groups. Markers developed to understand phylogeny and pathways of domestication could also serve in marker assisted breeding, a tool that would benefit the very difficult breeding process in *Musa.*

Most banana landraces are farmers' selection from intra- and interspecific hybrids of two diploid species, *Musa acuminata* Colla., donor of the A genome, and *Musa balbisiana* Colla., donor of the B genome (Simmonds, 1962). Most edible bananas are derived from these two diploid genomes, and are categorized into four groups (AAA, AAB, ABB and AAAB) according to the doses of the A and B genomes present (Figure 1). Natural polyploid cultivars include the triploid *M. acuminata* (AAA) sub-groups like Cavendish, Gros Michel and East African Highland banana groups, triploid hybrids (AAB) sub-groups which include the 'true' plantains, Pome and Silk types, and the triploid hybrid (ABB) cooking banana sub-groups. Synthetic tetraploid hybrids have also been produced in several genomic compositions (for example, AAAB and AABB) in international breeding programs (Escalant et al., 2002).

The East African Highland bananas (Mutika/Lujugira subgroup) belong to the AAA genome group, but are morphologically highly variable and are further classified as either beer or cooking varieties (Sebasigari, 1987; Karamura and Karamura, 1995). The AAA genome group (Figure 1) also contains the popular desert bananas, with 'Cavendish' clones and 'Gros Michel' being the most widely planted and consumed worldwide. Banana varieties with predominantly A genome (*M. acuminata*) produce sweet fruits (For example 'Giant Cavendish' - AAA genome), while those with high proportion of the B genome (*M. balbisiana*) produce starchy fruits (For example, 'Fougamou 1' - ABB genome). Tézenas du Montcel, (1988) identified seven subspecies of *M. acuminata* (namely, *microcarpa, malaccensis, burmannica, banksii, errans, burmannicoides,* and *truncata*). However, subsequent research using amplified fragment length polymorphism (AFLP) markers suggested that the seven subspecies may actually be only three subspecies: *microcarpa*, *malaccensis* and *burmannica* (Ude et al., 2002a, 2002b). Further, the subspecies *microcarpa* clustered very closely with the dessert bananas ('Gros Michel' and 'Yangambi km5'), indicating that the edible dessert bananas may have derived their A genome(s) from this taxon (Ude et al., 2002a; 2002b). Similarly, two genetic clusters were identified in *M. balbisiana* implying that the B genome is also genetically variable and that variants may be playing different roles in edible polyploid bananas (Ude et al., 2002b).

Shepherd (1988) reported that the AAB and ABB cultivars have different B genomes that arose naturally at different periods from unique parental genotypes in diverse geographical areas. However, he was not convinced that *M. balbisiana* played any role in the

Table 1. Cultivar name, identifier, group and genotype for the six *Musa* spp. germplasm accessions used in the parallel sequencing studies.

Accession name	Identifier[1]	Group[2]	Genotype
Calcutta 4	TARS 18242	subsp. *burmanicodies*	AA
Pisang Lilin	TARS 18252	subsp. *malaccensis*	AA
Tani	TARS 18046	*M. balbisiana*	BB
Manzano	TARS 17136	Silk	AAB
Obino l'Ewai	TARS 18239	Plantain	AAB
Fougamou 1	TARS 18022	Pisang Awak	ABB

[1]Germplasm maintained at the USDA-ARS Tropical Agriculture Research Station (TARS) and accessions can be found using the Germplasm Resources Information Network (GRIN). [2] Group in this case means cultivated subgroup or subspecies.

evolution of hybrid group cultivars. The AAB genome group in *Musa* consists of natural hybrids possessing 22 'A' and 11 'B' chromosomes. There is wide diversity in pulp characteristics and end use of the fruits of different AAB cultivars. 'Pome' and 'Silk' subgroups are the Indian and Brazil desert banana varieties that produce sub-acid, sweet fruits which are consumed fresh (Ude et al., 2002b). However, the AAB plantain subgroup produces long angular fruits with starchy pulp, which is not palatable unless cooked. Among triploids, the ABB are thought to have risen through artificial hybridization between *M. acuminata* and *M. balbisiana* followed by allopolyplodization (For example, 'Fougamou 1' and 'Bluggoe'). Such subdivision within the diverse polyploids may reflect differing contributions of specific subspecies in the phylogeny of individual groups of cultivars as previously suggested (Lebot et al., 1993).

Other plants with similar genomic structure to bananas, such as *Brassica* and cotton, are considered model systems for molecular phylogeny and marker-assisted selection (Iniguez-Luy et al., 2009; Qureshi et al., 2004). With the establishment of the Global Musa Genomics Consortium (*http://www.musagenomics.org/*) in the last century and the more recent reference genome reporting for *Musa* A (D'Hont et al., 2012) and B (Davey et al., 2013) species, researchers are using advanced technologies to study the crop.

Sequence-based characterization techniques are being used to study specific groups of banana cultivars to further understand their genetic structure and their contribution to the observed diversity of morphological and phenotypic characteristics that distinguish the different polyploid sub-groups and clones.

In the present project, a parallel sequencing study targeting primarily exons of nuclear genes was conducted in a subpanel of edible *Musa* cultivars. This was carried out in an effort to strategically reduce genome complexity in this polyploid species and score for DNA sequence variants (SNPs and INDELS) in the accessible genome. The long term goal of this study is to confirm diversity sampling between A and B genomes in *Musa* over a broad range of accessions and explore whether a particular sequence variant falls within or near a gene of interest that influence any phenotypic trait. Also this study will facilitate modern banana improvement, whereby breeders could benefit from an understanding of the phylogenic ancestry of the crop as well as pathways of domestication of the major cultivar groups.

MATERIALS AND METHODS

Plant collection

The materials for this study were obtained from the USDA-ARS Tropical Agriculture Research Station, in Mayaguez, Puerto Rico. They represented the global gene pool of commercial banana, with emphasis on cultivars with high value in breeding and utilization. DNA was extracted from two accessions harboring pure *M. acuminata* AA diploid genomes ('Calcutta 4' representing the *burmannicoides* subsp., and 'Pisang Lilin' representing the *malaccensis* subsp.), one accession harboring pure *M. balbisiana* BB diploid genome ('Tani'), two accessions representing the triploid hybrid AAB genome ('Manzano' from the Silk sub-group and 'Obino l'Ewai' from the Plantain sub-group), and one accession representing the triploid hybrid ABB genome ('Fougamou 1' of the Pisang Awak sub-group) (Table 1).

DNA extraction and fragmentation

Young trifoliate leaves were harvested and immediately frozen in liquid nitrogen. Frozen samples were ground with a mortar and pestle in liquid nitrogen. DNA was extracted from ~100 mg of powdered tissue using the DNeasy Plant Mini Kit (Qiagen, Alameda, CA) according to the manufacturer's instruction (including an RNase degradation step). A NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Waltham, MA) was used to quantify DNA before subjecting to Solution Hybrid Selection (SHS) process. Each DNA sample was treated with NEBNext® dsDNA Fragmentase (New England Biolabs, Ipswich, MA). Fragmentation of DNA to an average size of about 300 bp was verified using Bioanalyzer High Sensitivity DNA Kits (Agilent Technologies, Santa Clara, CA).

Library preparation

The DNA samples were subjected to library preparation with insert sizes of ~300 bp for paired-end sequencing. DNA sequencing

Table 2. *Musa* spp. subset of representative genomic targets in with their chromosomal location for NimbleGen baits library development.

Locus ID	Chr.	Start	End	Function
GSMUA_Achr1P06140_001	chr1	4945027	4948846	Granule-bound starch synthase 1, chloroplastic/amyloplastic~ WAXY
GSMUA_Achr3P08340_001	chr3	5924248	5935167	Phytochrome B~ PHYB
GSMUA_Achr11P07670_001	chr11	5942667	5944644	Alcohol dehydrogenase 1~ ADH1
GSMUA_Achr6P16390_001	chr6	10918532	10920408	Floricaula/leafy homolog~ FL
GSMUA_Achr7P27540_001	chr7	28530932	28535471	Homeobox protein KNOX3~ KNOX3
GSMUA_Achr2P23000_001	chr2	21865341	21865595	MYB family transcription factor, putative, expressed~ DNAJC2
GSMUA_Achr4P00320_001	chr4	266419	267033	bZIP transcription factor domain containing protein, expressed~ bzpF
GSMUA_Achr6P29060_001	chr6	29366717	29372486	whirly transcription factor domain containing protein, expressed~ rexB
GSMUA_Achr9P06750_001	chr9	4297866	4299193	DNA-binding WRKY domain-containing protein~ WRKY71
GSMUA_Achr7P26700_001	chr7	27942477	27949823	MADS-box transcription factor 1~ MADS1
GSMUA_Achr3P15930_001	chr3	16897534	16914480	MADS-box protein CMB1~ CMB1
GSMUA_Achr2P13710_001	chr2	15734217	15735431	Putative Transcriptional regulator STERILE APETALA~ SAP
GSMUA_Achr5P28610_001	chr5	28709554	28714790	ADP-ribosylation factor GTPase-activating protein AGD7~ AGD7
GSMUA_Achr11P01660_001	chr11	1124949	1126595	RAN GTPase-activating protein 1~ RANGAP1
GSMUA_Achr10P28690_001	chr10	31422716	31424264	stress-induced protein, putative, expressed~ TIF32
GSMUA_Achr2P12750_001	chr2	15109260	15109769	Zinc finger A20 and AN1 domain-containing stress-associated protein 9~ SAP9
GSMUA_Achr8P01310_001	chr8	1107887	1108219	EARLY flowering protein, putative, expressed~ aroK
GSMUA_Achr9P08660_001	chr9	5590379	5593943	plant neutral invertase domain containing protein, expressed~ dapD
GSMUA_Achr8P22470_001	chr8	27039690	27048032	Sucrose synthase 2~ SUS2
GSMUA_Achr10P18850_001	chr10	25416529	25419440	Sucrose transport protein SUT1~ SUT1
GSMUA_Achr8P05340_001	chr8	3475343	3489014	Cellulose synthase-like protein G3~ CSLG3
GSMUA_Achr7P19410_001	chr7	22227499	22231808	Cellulose synthase A catalytic subunit 9 [UDP-forming]~ CESA9
GSMUA_Achr4P11310_001	chr4	8106521	8106880	Dehydrin Xero 1~ XERO1
GSMUA_Achr5P15820_001	chr5	12318136	12322339	Catalase isozyme 2~ CAT2
GSMUA_Achr2P05850_001	chr2	10788052	10791226	Lipoxygenase A~ LOX1.1

libraries were prepared using Kapa Library Preparation Kit reagents and protocol (Kapa Biosystems, Wilmington, MA). This was carried out with end repair of the fragmented DNA followed by A-base addition to the blunt ends of each strand, and finally adapter ligation using Illumina TruSeq Adapters, (Illumina, San Diego, CA). Each adapter had a 'T'-base overhang on the 3'-end, providing a complementary site for ligating to the A-tailed fragmented DNA. The final recovery product was used as template in the pre-hybridization library amplification for enrichment and clean-up for subsequent steps. The library was subjected to electrophoretic evaluation in an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) using a DNA 1000 chip.

Probe design

Five hundred genes distributed throughout the 11 chromosomes were targeted for enrichment, based on the most retained gene ontology categories reported for the draft *Musa* sequence (D'Hont et al., 2012). Probes included classical single copy genes (for example, *Adh1, PhyB, Lfy* and *Waxy*) and members of several transcription factor families (For example, *MYB, MADS*-box factors and *WRKY*) which were found to be over-retained after the whole-genome duplication in *Musa* α/β (banana) (D'Hont et al., 2012) (Table 2).

The length of final targets designed ranged from ~300 to 6,000 bp with a mean length of 1,500 bp and a total length of approximately ~500 kb. Hybridization probe-sets for selected targets (average size 100 bp) are designed from consensus sequences calculated from the reference 'DH Pahang'-A genome (D'Hont et al., 2012) and by Roche NimbleGen (NimbleGen Systems, Madison, WI) using proprietary software algorithms with generalized parameters for probe sequence, hybridization temperature and length. The design of the probes for each of the 500 pairs of targets is aimed to overlap >80% of the targeted sequences (*Musa* reference genome A sequence).

Hybridization and MiSeq processing

Streptavidin-bead capture hybridization between the indexed libraries or template DNA and the biotinylated exon-derived probes was performed as previously published (Porreca et al., 2007; Gnirke et al., 2009). This was followed with post hybridization amplification (via Illumina adapters), purification of amplified samples and as before DNA quality check with Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). The resulting post capture enriched libraries with 150 bp paired-end processing was subjected to sequencing on MiSeq (Illumina, San Diego, CA) performed at the Ambry Genetics facility, Aliso Viejo, CA. Normalization was conducted to ensure that an even read coverage across samples being sequenced simultaneously was achieved. For this, one lane of Illumina sequencing was performed to determine the number of sequence-able molecules per library. MiSeq data was analyzed using RTA software ver. 1.13 and corresponding read lengths is expected to be 2 × 150 bases. Data was further processed using the Picard data-processing pipeline to generate BAM files. Alignment was performed using Burrows-Wheeler Aligner (BWA) software version 0.5.9 (Durbin and Li, 2009).

Table 3. Coverage Statistics of NimbleGen probes.

Parameter	Value
Offset in bp	100
Consolidated/Padded Regions	25
Final Target Bases	127,077
Predicted Coverage	113,896
% Target Bases Covered	89.6
Target Bases Not Covered	13,181
% Target Bases Not Covered	10.4

Sequence alignment

Filtered sequence reads were aligned to the annotated *Musa* reference genome for *M. acuminata* 'Double Haploid-Pahang' [Double Haploid-Pahang is a man-made diploid or double haploid *M. acuminata* (AA) plant that was used for genome sequencing; *http://banana-genome.cirad.fr/download/musa_pseudochromosome.fna.gz*] using the BWA alignment tool, allowing a maximum of four mismatches and one gap of up to 3 bp. The Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) was then used to query reads that cannot be aligned by BWA, first against the *Musa* reference genome with an e-value cutoff of 1e22 and then against the National Center for Biotechnology Information (NCBI) database using default settings. Intron/exon boundaries were further verified using the European Molecular Biology Laboratory (EMBL) program Gene Wise [*http://www.ebi.ac.uk/Tools/psa/genewise/*] and the GBrowse tools under the Banana Genome Hub (*http://banana-genome.cirad.fr/home*). Different minimum overlap identity rate ~90% was also tested to facilitate measuring on- and off-target rates for each library and each experiment and avoid comparison of putative close paralogs.

Variant detection

Following sequencing, samples were processed using CASAVA 1.8.2 (Illumina, San Diego, CA). This processing included demultiplexing, aligning 100 bp short reads to the reference sequence obtained from the Banana Genome Hub (http://banana-genome.cirad.fr/download/musa_pseudochromosome.fna.gz), and variant calling. Variant calls were made for Qsnp scores > 20 and when coverage > 10. The variant calls from CASAVA-based pipeline were not suitable, on their own, for this analysis of multi-nucleotide polymorphisms. To resolve this, sequence variants were scored among aligned reads for covered regions, using the FreeBayes polymorphism discovery algorithm (Garrison and Gabor, 2012), restricting the calls to those originally made by the CASAVA pipeline. All samples were then combined into a single VCF file using the VCFtools program vcf-merge. Subsequently, this combined VCF file was split using VCFtools such that the coding sequence (CDS) and non-CDS regions were in separate files. The CDS regions were defined by files downloaded from the Banana Genome Hub website for *M. acuminata* 'DH-Pahang' v1. The VCF file was then searched for variants that were classified as 'TYPE=mnp' by Freebayes. Sequence variants included binary SNPs and small indels, as well as allelic series of tri-SNPs and tetra-SNPs, multi-nucleotide polymorphisms (MNPs), and INDELS with a variable number of (repetitive) nucleotides. Reads marked as duplicate, with more than seven base mismatches, more than three separate gaps, or with mapping quality (MQ) <30 were excluded for variant calling. Sequence variants adjacent to indels, which may arise due to local misalignment were filtered using Genome Analysis ToolKit or GATK (McKenna et al., 2010). Finally, each variant was annotated using snpEff (version 4.0d, with a custom-build database) to identify variants that were missense/nonsense relative to the reference sequence.

RESULTS

Sequencing and alignment

An in-solution hybridization capture library targeting primarily exons of nuclear coding genes was designed. Baits targeting genomic sequences in the enrichment library were distributed across all 11 *Musa* chromosomes (Table 2). Genomic libraries from the six *Musa* cultivars were indexed with Illumina Truseq adapters. The samples were captured in groups of three into two pools and paired-end sequenced. In total, 14,779,498 read-pairs were obtained, representing ~45 Megabases of sequencing data. The cultivar-specific sequence index could be identified in 95-98% of the read-pairs (Supplementary Table 1).

Genome space coverage

As a consequence of the enrichment method, sequences aligned not only to target regions, but also to flanking and off-target regions. After the reads were mapped to the reference sequence, there were 2,197,401 bp (~2.2 Mb) sequenced, of which 545,559 (~0.5 Mb) was in coding regions (CDS). There were 3,284 CDS regions sequenced.

Almost all genomic regions and genes targeted by the enrichment library fell within the accessible regions as shown in Table 3.

Variant detection

Variant calls were made for Qsnp scores > 20 and when coverage is > 10. The chromosome and location of the variant are listed as well as the reference and alternate alleles and the estimated genotype. Genotypes are reported as follows:

(i) HOM_REF both alleles match the reference sequence. *This will not appear in variant reports because such reports only contain variants.
(ii) HOM_ALT both alleles are the alternate allele.
(iii) HET_REF one allele is the reference and one is the alternate allele
(iv) HET_ALT both alleles are different alternate alleles

Also given is whether the variant is located in a coding region (CDS) or non-CDS. Again SNPs could be annotated relative to the reference sequence and identified as 8,831 synonymous variants, 6,631 missense variants and 221 splice variants across the six accessions.

Table 4. Overview of DNA variants observed across the six accessions in the available genome.

Variant type	Sequenced genome (2,197,401 bp)	Non-CDS[1] (1,651,842 bp)	CDS[1] (545,559 bp)
Dinucleotide SNPs	1,850	1,641	209
Trinucleotide SNPs	6	6	0
Tetranucleotide SNPs	6	4	2
Other MNPs[2]	8	8	0
INDELS	11,672	11,203	469

[1]CDS = Coding regions. [2] MNPs = Multi-nucleotide polymorphisms.

Sequence diversity analysis

Table 4 shows a total of 13,542 putative sequence variants (SNPs, MNPs and INDELS) that were identified in the accessible genome (12,862 in non-coding and 680 in the coding). The density of substitution variants (SNPs and MNPs) was 7.8 times higher in non-coding regions than in coding regions, and the INDEL density was 24 times higher in non-coding regions. Across all cultivars, an average variant density of 1/802 bp in coding regions and 1/128 bp in non-coding regions was observed.

DISCUSSION

DNA variants such as single nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs), and insertions and deletions (INDELS) are different at the nucleotide sequence level among individuals or alleles and represent the basic units of genetic diversity (Uitdewilligen et al., 2013). Characterizing this diversity in polyploids via genotyping of sequence variants can be achieved by direct Sanger sequencing of PCR amplicons (Rickert et al., 2002; Sattarzadeh et al., 2006). However, the procedure is laborious and time-consuming, requiring unique primer sets to be designed in order to obtain uniform amplification parameters across alleles. Further, not only can the analysis not exceed a certain number of target genes, it is more expensive on a per-sequence-generated basis. Another high-throughput method to screen for polymorphic markers in different species is restriction-site-associated DNA (RAD) sequencing (Baird et al., 2008), which generates markers associated with designated genomic restriction sites. However, this method has several drawbacks. First, the nucleotide variants in the restriction site may interfere with digestion and cause null alleles in addition, the technique cannot target specific regions of interest or reduce genomic complexity.

Sequence capture is widely used for isolating targeted alleles from the background of an entire genome. The scale of the capture can range from hundreds to thousands of loci simultaneously (Metzker, 2005; Craig et al., 2008). By reducing the sequencing space per sample, this method makes multiplexing feasible, thereby reducing the overall sequencing cost. Secondly, it targets only a portion of the genome that is either necessary (for example, where a specific number of genes is required for adequate genomic coverage) or informative (for example, genes of a specific pathway) for the biological question being addressed, thereby reducing the complexity of the analysis (Grover et al, 2012). Lastly, given the redundancy of plant genomes, which typically include myriad gene duplications (Van de Peer et al., 2009; Jiao et al., 2011), and the fact that polyploidy is typical for many plants (Jiao et al., 2011), the read depth afforded by targeted NGS increases the possibility of identifying both the precise region (orthologs) of interest and its paralogs. The usual technical approaches under sequence capture involve hybridization of samples either to solid platforms/arrays or to solution-based, pooled oligonucleotide- or RNA-baits, both of which are complementary to the targeted genes (Davey et al., 2011). Probe design and synthesis may be outsourced to commercial service providers like NimbleGen and Agilent, who also provide a streamlined protocol for sequence capture suitable for laboratory use. Sequence capture is also advantageous in that *a priori* sequence information is required for only one taxon and subsequent capture baits designed to enrich the sequencing library may be used across taxa (at low levels of divergence).

Hybridization-based enrichment in plant research has increased in recent years. Fu et al. (2010) used sequential on-array hybridization to deplete repetitive elements from *Zea mays* genomic libraries, and then enrich the libraries for unique target loci. Saintenac et al. (2011) used solution hybridization to target nearly 3,500 dispersed loci (3.5 Mbp) from the larger genomes of allotetraploid wheats (*Triticum dicoccoides* and *T. durum* cv. Langdon, each nearly 10 Gb/1C) that were barcoded, pooled, and hybridized in a single reaction. Recently, Salmon et al. (2012) used a similar used a similar micro-array based hybrization capture to successfully target 500 pairs (homeologs) of selected genes in wild and domesticated *Gossypium hirsutum*, encompassing 550 kb of haploid transcript space (Salmon et al., 2012).

The recently sequenced genome of *M. acuminata* downloaded from the banana genome hub [*http://banana-*

genome.cirad.fr/download/musa_pseudochromosome.fna.gz] is appropriate for performing this next-generation parallel sequencing in banana. A solution-based sequence capture method to target *Musa* genomes for identifying native gene variation within the different subspecies of the *M. acuminata* as well as the *M. balbisiana* is reported for the first time in this work. Target enrichment allowed the achievement of sufficient sequence coverage depth across cultivars in our experiments. Importantly, bait hybridization enrichment allowed exclusion of repetitive regions of the *Musa* genome. The re-sequencing data from this method was instrumental to identify sequence variants as potential genomic markers in a non-model crop plant. We focused on single copy genes, for example from the set of conserved orthologous sequence genes (For example, *Lfy, Waxyn* and *WRKY*) and used uniquely mapped genes to define the genomic target regions.

To reduce sequencing costs, we multiplexed the different cultivars in a sequencing pool with custom index adapters. No index-specific bias was observed in the read counts and ~ 96% of the generated reads could be assigned to cultivars. We found consistent enrichment across all indexed samples, and virtually all target sequences were covered at a sufficient depth.

We used a mapping approach to align the sequencing data to the *Musa* reference genome sequence (D'Hont et al., 2012). Our observations led to detection of synonymous as well as missense variants distributed across the genome. An advantage of alignment to an annotated reference genome is that it allows prediction of whether a sequence variant falls within or near a gene of interest, and whether it is expected to cause a functional change in the protein product (For example, synonymous versus non-synonymous) that might alter the enzyme activity of the protein (Uitdewilligen et al., 2013). This can be very useful in determining whether a particular sequence variant is likely to be responsible for a phenotype of interest. A disadvantage of mapping sequence reads toward a reference sequence is that structural variation like chromosome rearrangements, inversions and large (transposon) insertions are likely to be missed. This can be partly avoided by *de novo* assembly, but computational difficulties associated with the assembly of highly diverse polyploid species like *Musa* makes mapping sequence reads to a reference sequence a more straightforward approach.

This study is the first to show high sequence diversity of *Musa* on a genome-wide scale even with a small number of cultivars. Our future endeavors aim to include a much larger number of *Musa* cultivars which will further increase the overall frequency of one variant for fewer number of bp in the cultivar population, both in non-coding and coding regions. There is a drawback with the presence of a limited draft genome and incomplete characterization for some genes. This may give rise to ambiguity in concluding whether any given gene is single copy as the result of random loss or selection. Our future efforts will also include functional characterization, phenotypic evaluations and population genetic studies of these genes as well as expanding the targeted regions further to span more intronic sequences to undermine their roles in controlling the morphological traits.

The resulting marker dataset is most useful for describing allele frequencies and nucleotide diversity. Also, our approach is an efficient means of producing data for the design of both high and low-density SNP genotyping assays applicable to a wide range of *Musa* cultivars, and the resulting tools can be used to address questions in population genetics and marker-trait association research. Thus, we report an excellent platform of using sequence variants as molecular markers which should pave the way for recognizing associations between sequence polymorphism and phenotypic variation in a polyploid species like *Musa*.

Conflict of Interests

The author(s) have not declared any conflict of interest.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Baird NA, Etter PD, Atwood, TS, Cyrrey MC, Shiver AL, Lewis ZA, Selker EU, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ (2008). Identification of genetic variants using barcoded multiplexed sequencing. Nat. Methods 5:887-893.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci AM, Weissenbach J, Ruiz M, Glaszmann JC, Quétier F, Yahiaoui N, Wincker P (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488(7410):213-217.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499-510.

Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J (2013). A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. BMC Genomics 5:683.

Durbin R, Li H (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Escalant JV, Sharrock S, Frison E (2002). The genetic improvement of *Musa* using conventional breeding, and modern tools of molecular and cellular biology. IPGRI, Rome, Italy. p. 17.

FAOSTAT (2014). http://faostat.fao.org/site/291/default.aspx (verified Sep. 2014). FAO, Rome.

Fu Y , Springer NM, Gerhardt DJ, Ying K, Yeh CT, Wu W, Swanson-Wagner R, D'Ascenzo M, Millard T, Freeberg L, Aoyama N, Kitzman J, Burgess D, Richmond T, Albert TJ, Barbazuk WB, Jeddeloh JA,

Schnable PS (2010) Repeat subtraction-mediated sequence capture from a complex genome. Plant J. 62: 898-909 .
Garrison EM Gabor M (2012). Haplotype-based variant detection from short-read sequencing. ArXiv e-prints: 1207.3907.
Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27:182-189.
Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99(2):312-9.
Iniguez-Luy FL, Lukens L, Farnham MW, Amasino RM, Osborn TC (2009) Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*. Theor. Appl. Genet. 120:31-43.
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, de Pamphilis CW (2011). Ancestral polyploidy in seed plants and angiosperms. Nature 473(7345):97-100.
Karamura DA, Karamura EB (1994). A provisional checklist of bananas in Uganda. INIBAP, Montpellier, France.
Lebot V, Meilleur AB, Manshardt RM (1994). Genetic diversity in Eastern Polynesian Eumusa bananas. Pac. Sci. 48:16-31.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.
Metzker ML (2005). Emerging technologies in DNA sequencing. Genome Res. 15: 1767-1776.
Qureshi SN, Saha S, Kantety RV, Jenkins JN: EST-SSR (2004). A new class of genetic markers in cotton. J. Cotton Sci. 8:112-123
Rickert AM, Premstaller A, Gebhardt C, Oefner PJ (2002). Genotyping of SNPs in a polyploid genome by pyrosequencing (TM). Biotechniques 32:592-593.
Saintenac C, Jiang D, Akhunov E (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12: R88 .
Salmon A, Udall JA, Jeddeloh JA, Wendel J (2012) Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. G3 (Bethesda) 2(8):921-30.
Sattarzadeh A, Achenbach U, Lübeck J, Strahwald J, Tacke E, Hofferbert HR, Rothsteyn T, Gebhardt C (2006). Single nucleotide polymorphism (SNP) genotyping as basis for developing a PCR-based marker highly diagnostic for potato varieties with high resistance to *Globodera pallida* pathotype Pa2/3. Mol. Breed. 18:301-312.
Sebasigari K (1987). Morphological taxonomy of *Musa* in Eastern Africa. In: Persley, G.J. and De Langhe, E.A. (eds.) Banana and Plantain Breeding Strategies, ACIAR Proceedings 21, ACIAR Canberra. pp. 172-176.
Shepherd K (1988). Observation on *Musa* taxonomy. In: Identification of genetic diversity in the genus *Musa*. Proc. Int. Workshop, Los Baños, Philippines. INIBAP, Montpellier, France. pp. 158-165.
Simmonds NW (1962). Evolution of the bananas. London: Longmans, Green and Co.
Tézenas du Montcel H (1988) *Musa acuminata* subspecies *banksii*: status and diversity. Paper presented at the INIBAP/PCARRD Workshop on the identification of genetic diversity on the identification of genetic diversity in the genus *Musa*, Los Banos, Phillipines, 5-9 Sep.
Ude G, Pillay M, Nwakanma D, Tenkouano A (2002a). Analysis of genetic diversity and sectional relationships in *Musa* using AFLP markers. Theor. Appl. Genet. 104:1239-1245.
Ude G, Pillay M, Nwakanma D, Tenkouano A (2002b). Genetic diversity in *Musa acuminata* Colla. and *Musa balbisiana* Colla. and some of their natural hybrids using AFLP markers. Theor. Appl. Genet. 104:1246-1252.
Uitdewilligen JG, Wolters AM, D'hoop BB, Borm TJ, Visser RG, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One 8(5):e62355.

Supplementary Table 1. Cultivar-specific sequence index identified in 95-98% of the read-pairs.

Accession	Group	Read-pairs	Mb	% Perfect indices
Manzano	Silk	3491071	1047	99.16
Pisang Lilin	subsp. *malaccensis*	847170	254	95.42
Tani	*M. balbisiana*	2907131	872	99.05
Calcutta 4	subsp. *burmanicodies*	1717363	515	98.77
Obin l'Ewai	Plantain	2979951	894	98.76
Fougamou 1	Pisang Awak	2836812	851	98.66